# Spectral clustering and idea of using clustering algorithms in genetics

Max Planck Institute of Psychiatry
German Research Institute of Psychiatry

Helena Pelin, MPI of Psychiatry Munich

Topic of my Master Thesis on Faculty of Science Zagreb, Croatia

## 1 Introduction

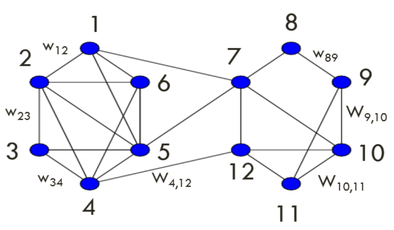
Clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, biology to social sciences or psychology. Clustering belongs to the *unsupervised machine learning techniques* whose aim is to describe hidden structure of the "unlabelled" data, meaning that a classification or categorization of the data points is not included in the observations. In this process of describing the hidden structure of the data, clustering has a task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups (clusters). There are different types of clustering methods and here we are going to cover Spectral Clustering method which is graph theoretic type of clustering, meaning that it treats clustering task as a graph partitioning problem. For this purposes, one has to define the so called Affinity matrix which contains values of some similarity measure between our data points and find eigenvectors of the corresponding Laplacian matrix of the graph.

## 2 Clustering as graph partitioning

**Aim of the clustering method** in general is to achieve the high similarity within objects in the same group and at the same time weak similarity between objects that are in different groups

Spectral Clustering treats the clustering task as a graph partitioning problem and therefore, we need to present our data in a convenient way – data are represented in a form of a weighted undirected graph where the nodes are the points in a feature space and the edges represent similarities between the points:

- We need to cluster n objects into groups
- We define G = (V,E), where V is a set of nodes $v_1,..v_n$ and E is a set of edge weights $w_{ij}$ which represent similarity between object i and j
- Graph has affinity matrix W and Laplacian matrix L:




$$W = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & a_{n,2} & \cdots & w_{n,n} \end{pmatrix} \qquad D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$$

$$d_i = \Sigma_{j=1}^{n} w_{ij}$$

$$L = D - W$$

## 3 Partitioning algorithm

Input: Affinity matrix W and number of clusters k:

(S1) Compute the Laplacian matrix L

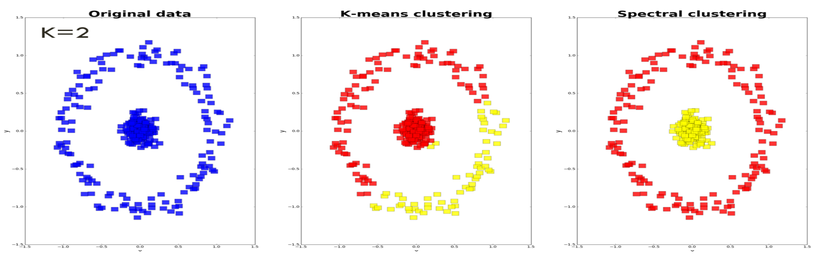
(S2) Compute the first k eigenvectors $\{v_1, v_2, \ldots, v_k\}$ of the **generalized eigenproblem** $Lv = \lambda Dv$.

(S3) Let $V \in R^{(n \times k)}$ be the matrix containing vectors $\{v_1, v_2, \ldots, v_k\}$ as columns

(S4) For i = 1,…,n, let $y_i \in R^k$ be the vector corresponding to the i-th row of matrix V

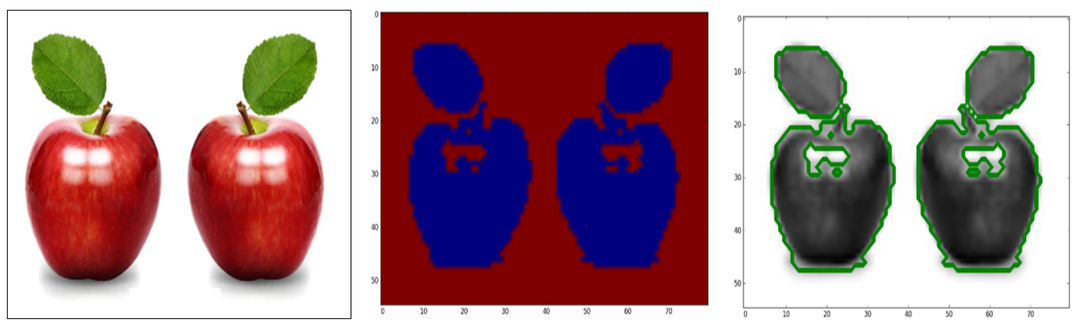
(S5) Cluster the points $(y_i)_{i=1,\ldots,n}$ in $R^k$ with the **k-means** algorithm into clusters C1, . . . , Ck.

Output: Clusters $A_1, \ldots, A_k$ with $A_i = \{j | y_j \in C_i\}$

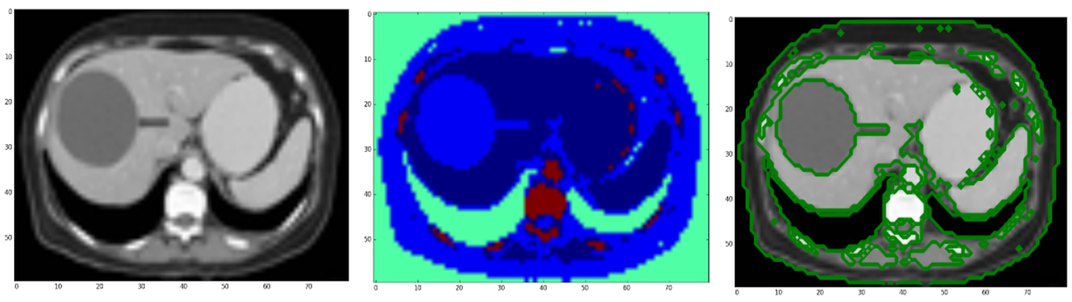



**Spectral Clustering v.s. K-means clustering**
Nr. of clusters = 2

## 4 Examples



**Dividing objects from the background**
Nr. of clusters = 2



**Detecting objects (e.g. a cyst) on the image**
Nr. of clusters = 4

## 5 Clustering in my PhD project

My PhD project is a part of the FOR2107 (Forschergruppe) by the Deutsche Forschungsgemeinschaft whose aim is, among others, to produce de-novo subgrouping of patients based on biological, cognitive and psychopathological longitudinal data.

I would focus on developing methods for finding clusters of patients with different affective disorders and controls in high-dimensional data spaces, based on their genetic and other omics-information and also use and train the algorithms which would create a predictive framework for automated medical diagnosis.

**FOR2107**

**Genotype data**
- Illumina PsychChip
- 2.500 participants:
  - ~ 1000 healthy controls
  - ~ 1500 patients mostly with Major depressive disorder and Bipolar Disorder, some with Schizophrenia and Schizoaffective Disorder

**Other available data:**
- Methylation data,
- Imaging data,
- Transcriptomics data,
- Clinical variables….

etc.

References:
[1] Shi, Malik – Normalized cut and image segmentation
[2] Stella X. Yu, Yianbo Shi – Multiclass spectral clustering
[3] Ulrike von Luxburg – Tutorial on spectral clustering
This work was the topic of my Master thesis on the Faculty of Science, Zagreb, Croatia